R-Bind: Unified Enhancement of Attribute and Relation Binding in Text-to-Image Diffusion Models

Huixuan Zhang **Xiaojun Wan**
Wangxuan Institute of Computer Technology, Peking University
zhanghuixuan@stu.pku.edu.cn,wanxiaojun@pku.edu.cn

Abstract

Text-to-image models frequently fail to achieve perfect alignment with textual prompts, particularly in maintaining proper semantic binding between semantic elements in the given prompt. Existing approaches typically require costly retraining or focus on only correctly generating the attributes of entities (entity-attribute binding), ignoring the cruciality of correctly generating the relations between entities (entity-relation-entity binding), resulting in unsatisfactory semantic binding performance. In this work, we propose a novel training-free method R-Bind that simultaneously improves both entity-attribute and entity-relation-entity binding. Our method introduces three inference-time optimization losses that adjust attention maps during generation. Comprehensive evaluations across multiple datasets demonstrate our approach's effectiveness, validity, and flexibility in enhancing semantic binding without additional training.

1 Introduction

Text-to-Image (T2I) models have achieved remarkable capabilities in synthesizing high-quality, photorealistic images (Betker et al., 2023; Esser et al., 2024). However, these models still face significant challenges in faithfully interpreting and following user prompts. Common failure modes include inaccuracies in object generation, attribute assignments, and relationships between entities (Li et al., 2024a), highlighting persistent limitations in semantic binding.

Numerous approaches have been proposed to address these limitations. Training-based methods such as GLIGEN (Li et al., 2023), CoMPaSS (Zhang et al., 2024) demonstrate promising results but face two critical challenges including high computational resource requirements and uncertain generalization capabilities across diverse scenarios.

Training-free approaches have also been explored to address these limitations. SynGen (Rassin et al., 2023) introduces specialized losses for entity-attribute binding (correctly generating the attributes of an entity, e.g., brown cat). Subsequent works like (Li et al., 2024b; Meral et al., 2024) further develop attention-based modifications for this purpose. However, these methods focus exclusively on entity-attribute binding, neglecting other crucial prompt semantics, making them unable to address many semantic binding problems. Notably, they fail to address entity-relation-entity binding (correctly generating relations between entities, e.g., a cat chasing a dog), which is equally (if not more) vital for faithful text-to-image generation.

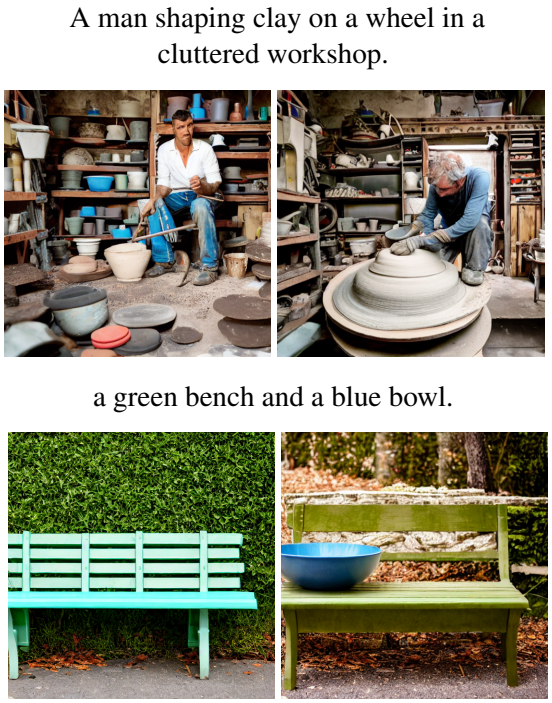

Figure 1: Examples of semantic binding using our method. The images on the left are the original generation results by SD-1.5, and the images on the right are generation results using SD-1.5 equipped with our method.

In this study, we propose a novel unified approach to enhance both entity-attribute and entity-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6856–6870

relation-entity binding in text-to-image generation by manipulating attention maps during inference. Our key innovation lies in establishing relation-aware attention patterns, where both entities maintain similar attention focus with this relation, while preserving distinct attention map between entities themselves to prevent entity confusion. Simultaneously, we enforce different image regions attend to distinct prompt components, preventing information omission or mixing. These principles with other observations are implemented through three carefully designed losses to perform inference-time optimization during denoising, effectively enforcing correct semantic bindings while penalizing incorrect bindings without requiring additional training. As illustrated in Figure 1, our method effectively handles complex prompts, with extensive experiments demonstrating its effectiveness across diverse scenarios and both U-Net and DiT-based diffusion architectures.

To summarize, our main contributions are listed as follows [1]:

- We introduce a novel semantic binding approach which can address both entity-attribute and entity-relation-entity semantic binding with three carefully designed losses.
- Our method is training-free and model-agnostic, effective in both U-Net based and DiT based diffusion models, making it widely available.
- Extensive experiments including both automatic evaluation and human study demonstrate the superiority of our method against baselines and comparison methods, with an average of 12.8% improvement on SD-1.5 against the strongest baseline.

2 Related Works

2.1 Diffusion Models

(Ho et al., 2020) first introduced DDPM, which serves as the foundation for subsequent diffusion models. In diffusion models, there are generally two types of conditioning algorithms: classifier guidance (Dhariwal and Nichol, 2021) and classifier-free guidance (Ho and Salimans, 2022). (Rombach et al., 2022) proposes conducting denoising in latent space, a technique that has proven highly successful.

[1]Our code is available at https://github.com/lleozhang/R-Bind.

Many studies (Podell et al., 2023; Esser et al., 2024; Chen et al., 2024; Ho et al., 2022; Peebles and Xie, 2023) present applicable text-to-image diffusion models using classifier-free guidance. Despite the success of them, current text-to-image diffusion models still suffer from failures in alignment with text prompts.

2.2 Improving Semantic Binding in Diffusion Models

Many previous works have discussed ways of improving semantic binding in diffusion models. GLIGEN (Li et al., 2023) utilizes grounded generation, while CoMPaSS (Zhang et al., 2024) proposed a specific module for spatial understanding. ELLA (Hu et al., 2024b) utilizes a large language model for better text understanding, and CoMat (Jiang et al., 2024) utilizes a segmentation model to enhance training. Ranni (Feng et al., 2024) and TokenCompose (Wang et al., 2024) are two additional methods. However, these methods are all training-based methods, which face the problem of high cost and a lack of generalization ability.

There are also training-free methods. Attent-and-Excite (Chefer et al., 2023) first proposes modifying attention map and increasing the attention score of entities. Divide-and-Bind (Li et al., 2024b) further proposes entity-attribute binding using attention map. SynGen (Rassin et al., 2023) and CONFORM (Meral et al., 2024) introduces negative loss to further facilitate semantic binding, while ToMe (Hu et al., 2024a) proposes token merging for entity-attribute binding. However, all of these methods consider only entity-attribute binding, with more complex scenarios containing relation unexplored, limiting their practicability.

3 Preliminaries

Despite the complexity of text-to-image diffusion models, generally a text-to-image diffusion model contains a denoising network (either U-Net or DiT) ϵ_θ and a noise scheduler F. Given a text prompt p, at each denoising step t, the denoising network ϵ_θ makes two predictions $\epsilon_\theta(x_t, t, c)$ and $\epsilon_\theta(x_t, t, \phi)$, where c is the text embedding of the given text prompt p and x_t is the noise map at timestep t. The prediction following classifier-free guidance is $z_t = \epsilon_\theta(x_t, t, \phi) + \tilde{w}(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \phi))$, where $\tilde{w}$ is a hyper-parameter namely guidance scale. Then, using noise scheduler F, we have $x_{t-1} = F(x_t, z_t, t)$. After a total of T denoising

steps, we reach the final denoising result x_0.

Inspired by previous works (Chefer et al., 2023), at a certain denoising step t, if we can find a loss function $\mathcal{L}$ which measures how well the generation process satisfies some constraints that probably indicate a good generation result, we can perform a gradient descent on x_t as:

$$x_t^{'} = x_t - \alpha \frac{\partial \mathcal{L}}{\partial x_t} \tag{1}$$

we can use $x_t^{'}$ in the following inference $\epsilon_\theta(x_t, t, c), \epsilon_\theta(x_t, t, \phi)$ to achieve $x_{t-1}^{'}$ instead of x_t for a better generation result.

Despite the various design choices of ϵ_θ, there are generally always cross-attention operations between the noise map x_t and the text embedding c to help condition on the given text. Formally, given a noise map $x_t \in \mathbb{R}^{C \times h \times w}$, text embedding $c \in \mathbb{R}^{L \times C^{'}}$, $C, C^{'}$ are corresponding feature dimensions, h, w are the height and width of the noise map, L is the length of text prompt. For a cross-attention layer with H attention heads, the corresponding attention map is $A^{(0)} \in \mathbb{R}^{H \times I \times L}, I = h \times w$. Suppose there are K attention layers, the final attention map is $A^{(1)} \in \mathbb{R}^{K \times H \times I \times L}$. We average the final attention map across different layers and heads for further discussion, which is $A \in \mathbb{R}^{I \times L}$

4 Our Method: R-Bind

4.1 Background and Motivation

Inspired by previous work (Chefer et al., 2023; Rassin et al., 2023), we similarly identify improper attention focus as a factor in failed semantic binding. However, while existing studies have exclusively addressed entity-attribute binding scenarios, the critical case of entity-relation-entity binding remains unexplored. To illustrate this failure in entity-relation-entity binding, consider a text prompt “a man on the left of a lamp", we visualize the average attention map of the relation part “on the left of" in Figure 2.

Our analysis reveals a critical phenomenon during denoising: while two distinct regions initially attend to the relational tokens (i.e., “on the left of"), this focus gradually collapses to a single region as denoising progresses. This directly leads to semantic binding failures, incorrectly positions “the man below the lamp" rather than “on the left of" it. This observation demonstrates that maintaining proper attention focus throughout the denoising process

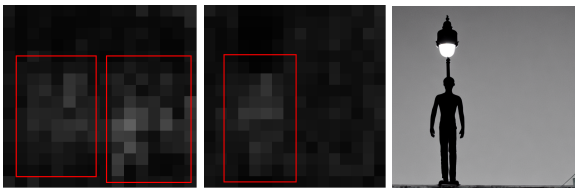

Figure 2: Example of a failure generation. The left shows the attention map at the first denoising step, the middle shows the attention map after 10 denoising steps, and the right shows the final generation result.

is essential for achieving correct semantic binding for entity-relation-entity binding, leading to our method R-Bind.

4.2 R-Bind

Our method R-Bind operates through two stages: semantic extraction and semantic binding enhancement. First, we automatically parse the input prompt to extract semantic information, including entities, attributes, and relations. We then apply three semantic binding losses using extracted semantic information to ensure proper semantic binding through inference-time optimization. The complete framework of our approach is illustrated in Figure 3.

4.2.1 Semantic Extraction

We consider a more generalized semantic binding in this work, including both entity-attribute binding and entity-relation-entity binding. The first step is to extract these semantics from the given prompt.

For a given prompt p comprising tokens $(t_1, ..., t_L)$, we categorize semantic components as follows: **entity tokens** are tokens directly representing objects, like “cat"“car". **Attribute tokens** are tokens describing entity properties without referencing other entities (e.g., "brown" in "a brown cat"). Note that in the prompt “a cat chasing a dog", “chasing a dog" is not viewed as an attribute, since it contains another entity. **Relation tokens** are tokens expressing inter-entity connections (e.g., "chasing" in "a cat chasing a dog"). Note that we consider all kinds of relations in this work instead of only spatial relations, further broadening applicability. For any certain entity, attribute, relation, there can be one or more tokens corresponding to it due to the complexity of expression or tokenization.

With these definitions, we can extract **Entity** set $S_e = (e_1, .., e_g)$, where e_i represents entity tokens determining one entity, like “cat"“dog"; **Entity-Attribute** set $S_{ea} = \{(e_1, a_1), ..., (e_n, a_n)\}$, where a_i represents attribute tokens describ-

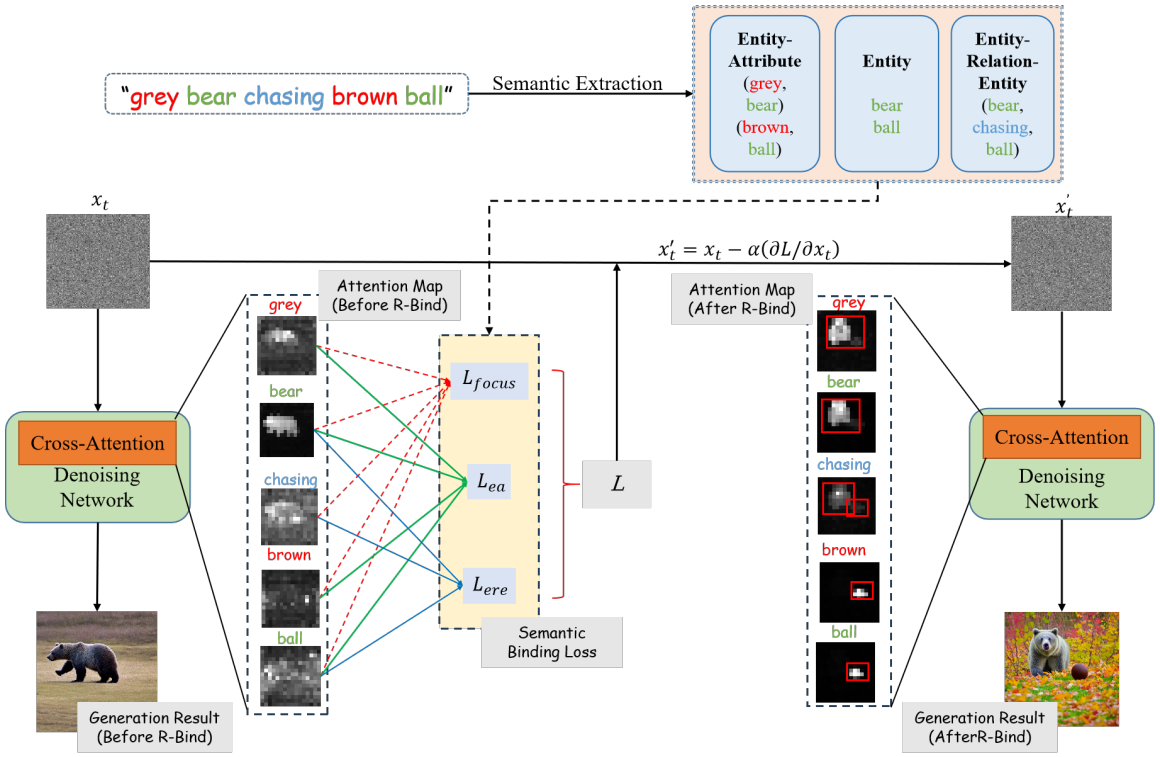

Figure 3: Overview of our method R-Bind. We use green in the texts to represent entity tokens, red to represent attribute tokens, and blue to represent relation tokens. Our method contains two steps: Semantic Extraction (as shown on the upper part) and Semantic Binding (as shown on the lower part). The left of the lower part shows the generation result of original model, while the right of the lower part shows the generation result of R-Bind. The middle part details how semantic binding is performed through inference-time optimization.

ing e_i, like e_i="cat" and a_i="brown" given "brown cat"; **Entity-Relation-Entity** set $S_{ere} = \{(e_1^1, r_1, e_1^2), ..., (e_m^1, r_m, e_m^2)\}$, where r_i represents the corresponding relation tokens describing the relation between e_i^1 and e_i^2. For example, given "cat chasing dog", we have e_i^1 being "cat", r_i being "chasing" and e_i^2 being "dog". The extraction of this semantic information can be performed using either a parser or an LLM.

4.2.2 Enhancing Semantic Binding

In the following description, we use D as a distance measure between two 1-d vectors, which in this work is selected as symmetric KL Divergence:

$$D(p, q) = \frac{1}{2}(D_{KL}(p||q) + D_{KL}(q||p)) \quad (2)$$

$$D_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} \quad (3)$$

For simplicity, we take the case that each e, a, or r corresponds to only a single token to illustrate our method (without loss of generality). For discussion about the case containing multiple tokens, please refer to Appendix A. We use $A[t] = A[:, t] \in \mathbb{R}^I$, which is a 1-$d$ vector representing the attention map of a certain token (t is a single token). For two tokens t_1, t_2, we note

$$D_t(t_1, t_2) = D(A[t_1], A[t_2]) \quad (4)$$

Focus Distribution *Focus Distribution* considers some basic principles that the attention map should follow. First of all, distinct positions in the noise map should attend to different parts in the prompt to prevent information mixing or omission. Positions farther apart in the noise map should exhibit greater divergence in their attention maps. For instance, as shown in Figure 3, the problematic overlap between attention regions for "ball" and "bear" leads to failed generation of the ball object. By strategically separating this attention focus, we achieve more accurate and reliable generation of all specified entities.

Secondly, each entity token should be focused by at least one position to avoid missing an entity. This is a similar observation with (Chefer et al., 2023).

For an attention map A, we note $\hat{A}[x] = A[x, :] \in \mathbb{R}^L$, which is a 1-$d$ vector representing the at-

tention map of a certain position in the noise map. Specifically, x corresponds to position (i, j) in the noise map where $x = i * w + j, i < h, j < w$. For each position in the noise map $x = i * w + j$, considering another position $y = p * w + q$, we can calculate their Manhattan Distance as $d(x, y) = |i - p| + |j - q|$. Therefore, we can construct a weight matrix $W \in \mathbb{R}^{I \times I}, W_{xy} = d(x, y)$ and $\tilde{W}$ is obtained by row-normalizing W. Therefore, denote $A_w = \tilde{W}A$, we can maximize the distance between A and A_w to achieve our goal of making farther positions in the noise map have different attention focus on the prompt. Combining the above analysis, the final *focus* loss is designed as:

$$\mathcal{L}_{focus} = -\frac{1}{I}\sum_{x} D(\hat{A}[x], \hat{A}_w[x]) - \min_{e \in S_e} \max_{x} A[x, e] \tag{5}$$

Entity-Attribute Binding Inspired by (Rassin et al., 2023), entity-attribute binding requires attention alignment between entity and attribute within an entity-attribute pair while maintaining separation between different pairs. Specifically, an attribute token (e.g., “grey" in Figure 3) should exhibit high attention similarity with its corresponding entity token (“bear"), while showing low attention similarity with unrelated entities (“ball"). This ensures visual attributes correctly bind to their target entities without interfering with other objects. We formalize this principle through our *entity-attribute binding* loss:

$$\mathcal{L}_{ea} = \sum_{(e_i, a_i)} [D_t(e_i, a_i) - \frac{1}{|Z|} \sum_{(e_j, a_j)} K((e_i, a_i), (e_j, a_j))] \tag{6}$$

where $|Z|$ is a normalizing factor, K is a measurement between two entity-attribute pairs. $K((e_i, a_i), (e_j, a_j)) = D_t(e_i, e_j) + D_t(e_i, a_j) + D_t(e_j, a_i) + D_t(a_i, a_j)$. To avoid separating potentially related information, we only calculate $K((e_i, a_i), (e_j, a_j))$ if and only if $e_i \neq e_j$ and $a_i \neq a_j$, otherwise $K((e_i, a_i), (e_j, a_j)) = 0$.

We would like to note that, a key distinction between our implementation and (Rassin et al., 2023) lies in the composition of the contrastive part. (Rassin et al., 2023) contrastive term includes all tokens not considered an attribute of the entity, whereas our formulation is more refined, only taking other token-relation pairs. This design choice is deliberate: it ensures that relation tokens related to an entity are not inadvertently included in the negative samples, which is crucial for accurately generating entity-relation-entity information.

Entity-Relation-Entity Binding Entity-relation-entity binding requires coordination of attention patterns across three components: two entities and their relation. The attention of relation tokens must align with both entities to properly generate this relation (e.g., "chasing" with both "bear" and "ball" in Figure 3), while the entities themselves must maintain distinct attention map to preserve their individual identities. This dual constraint ensures that the relationship is visually represented, and the entities remain clearly differentiated in the generated image. Also, attention of entities and relations within different triples should also be separated to avoid confused generation results.

Combining the objectives above, we achieve the *entity-relation-entity* loss as:

$$\mathcal{L}_{ere} = \sum_{(e_i^1, r_i, e_i^2)} [D_t(e_i^1, r_i) + D_t(e_i^2, r_i) - \min (D_t(e_i^1, e_i^2), \frac{1}{|Z|} \sum_{(e_j^1, r_j, e_j^2)} K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2)))] \tag{7}$$

Similarly, $|Z|$ is a normalizing factor and K is a distance measurement between two entity-relation-entity pairs. $K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2)) = D_t(e_i^1, r_j) + D_t(e_i^2, r_j) + D_t(e_j^1, r_i) + D_t(e_j^2, r_i) + D_t(e_i^1, e_j^1) + D_t(e_i^2, e_j^2)$.

We also calculate $K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2))$ if and only if $e_i^1 \neq e_j^1, r_i \neq r_j, e_i^2 \neq e_j^2$, otherwise $K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2)) = 0$.

Based on these above analysis, our final loss is:

$$\mathcal{L} = \mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere} \tag{8}$$

With our final loss (Equation 8), we can perform inference-time optimization with Equation 1. Details about our method design can be found in Appendix A.

5 Experiment Setup

5.1 Baseline Methods

To comprehensively evaluate our method, we implement it on two distinct base models: Stable-Diffusion-1.5 (SD-1.5) (Rombach et al., 2022) and Stable-Diffusion-3 (SD-3) (Esser et al., 2024), which differ in both architecture and capability. On SD-1.5, we compare against five training-free baselines: Attend-and-Excite (A&E) (Chefer et al.,

Model Base	Method Name	T2ICompBench (Color)	T2ICompBench (Spatial)	GenAIBench (Attribute)	GenAIBench (Spatial)
SD-1.5	Base model	37.6	8.7	63.4	62.0
	A&E	54.4	10.3	66.2	64.7
	SynGen	55.7	10.9	65.1	62.4
	ToMe	40.6	8.8	63.7	61.4
	D&B	55.3	10.4	64.5	61.7
	CONFORM	68.7	10.2	63.6	61.9
	R-Bind (Ours)	**68.8**	**15.6**	**68.2**	**67.9**
SD-3	Base Model	80.3	31.2	80.1	78.4
	R-Bind (Ours)	**82.5**	**32.0**	**80.7**	**79.4**

Table 1: Main Results of our method and compared baselines on test datasets. We use the evaluation metrics proposed corresponding to each test set, which means BLIP-VQA for T2ICompBench(Color), UniDet for T2ICompBench(Spatial), and VQAScore for GenAIBench.

2023), SynGen (Rassin et al., 2023), ToMe (Hu et al., 2024a), Divide-and-Bind (D&B) (Li et al., 2024b), and CONFORM (Meral et al., 2024). Notably, these baselines cannot be directly applied to SD-3 due to architectural differences, limiting their comparison to SD-1.5 only. For fair evaluation, we exclude all training-based methods from our comparisons.

5.2 Benchmarks and Metrics

We employ both constructed structured prompts and more natural prompts across multiple benchmarks. For constructed structured prompts, we utilize the color and spatial splits from T2I-CompBench (Huang et al., 2023), adopting their original metrics (BLIP-VQA and UniDet)(Huang et al., 2023). We also leverage GenAIBench (Li et al., 2024a), organizing its prompts into two test sets: GenAIBench(attribute) containing all prompts testing attribute binding skill, and GenAIBench(spatial) comprising prompts evaluating spatial relation skill. While these sets are not mutually exclusive and involve multiple skills, this categorization enables clearer analysis of specific capabilities. We employ VQAScore (Li et al., 2024a) for GenAIBench evaluation. We generate one image per prompt for evaluation.

5.3 Implementation Details of Our Method

We select the first 50% of total inference steps performing R-Bind following (Rassin et al., 2023), and perform gradient descent (Equation 1) twice each step. For SD-1.5, since it is a U-Net architecture and the resolution of attention map changes, we gather and average all attention maps at resolution 16×16 to calculate $\mathcal{L}$, still following (Rassin et al., 2023). For SD-3, since it uses a DiT architecture and the resolution of attention maps remains the same, we gather and average all cross attention maps. To maintain a fair comparison, we use the same noise prior for the same base model. The semantic extraction can be performed with any powerful LLMs, and we use Gemma-3 (Team, 2025) for semantic extraction (without the loss of generality). More details about our experiments and implementation can be found in Appendix B.

6 Experiment Results and Analysis

6.1 Main Results

The experimental results in Table 1 demonstrate R-Bind's superior performance across all datasets against all baseline when implemented on SD-1.5, directly supporting the effectiveness of R-Bind. We would also like to note that, while baseline methods are specifically designed for entity-attribute binding, they nevertheless show slight improvements over the base SD-1.5 model on the entity-relation-entity focused T2ICompBench(Spatial) dataset. We attribute this unexpected gain to their implicit enhancement of entity generation or treatment of relations as attributes. However, this implicit enhancement is not enough for performing correct entity-relation-entity binding, indicating that previous baselines are unable to address entity-relation-entity binding effectively.

CONFORM emerges as a strong competitor on T2ICompBench(Color), matching our method's performance on this entity-attribute focused dataset. However, its superiority diminishes on other datasets, revealing limitations in complex applications. In contrast, R-Bind maintains consistently high performance across all scenarios, demonstrating robust practical applicabiliy.

Notably, several baselines eve exhibit performance degradation on GenAIBench(Spatial), suggesting that over-optimization for entity-attribute

$\mathcal{L}_{focus}$	$\mathcal{L}_{ea}$	$\mathcal{L}_{ere}$	T2ICompBench (Color)	T2ICompBench (Spatial)	GenAIBench (Attribute)	GenAIBench (Spatial)
✗	✗	✗	37.6	8.7	63.4	62.0
✓	✗	✗	55.0	10.8	68.1	65.6
✗	✓	✗	64.4	8.7	65.9	63.5
✗	✗	✓	39.0	12.8	64.7	63.3
✓	✓	✗	68.2	10.8	**68.2**	66.1
✓	✗	✓	55.6	**15.6**	68.0	67.4
✗	✓	✓	64.4	12.8	65.7	65.1
✓	✓	✓	**68.8**	**15.6**	**68.2**	**67.9**

Table 2: Ablation study of our method using SD-1.5. ✓ refers to the corresponding loss is applied to the final loss $\mathcal{L}$, while ✗ indicates the loss is not applied to $\mathcal{L}$. The first line corresponds to the base model, and the last line corresponds our whole method R-Bind. The evaluation metrics remain the same as before.

binding may actually impair model performance in some scenarios. This finding underscores the importance of jointly addressing both entity-attribute and entity-relation-entity binding, as implemented in our approach.

SD-3 is built with different architecture with SD-1.5, with no prior work having explored semantic binding methods on this state-of-the-art model. Our results demonstrate that attention-based semantic binding remains effective even for SD-3's DiT architecture, with consistent performance gains across all datasets. These findings validate both the generalizability of our approach and its potential applicability to cutting-edge diffusion models. The observed improvements are further corroborated by our human evaluation study (Section 6.3), which provides additional evidence of the method's practical benefits.

6.2 Ablation Study

While the design of our three losses is intuitive, we conduct comprehensive ablation studies on SD-1.5 to rigorously evaluate each component's contribution. The results are shown in Table 2.

The ablation studies yield several key insights. First, any combination of $\mathcal{L}_{focus}, \mathcal{L}_{ea}, \mathcal{L}_{ere}$ produces better results than the base SD-1.5 model, with some combinations even matching or surpassing the baseline methods in Table 1. This confirms the effectiveness of each individual loss component. Second, we observe consistent performance gains when adding additional losses. For example, $\mathcal{L}_{focus} + \mathcal{L}_{ea}$ outperforms $\mathcal{L}_{focus}$ alone, and the full combination $\mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere}$ achieves the best results. This observation clearly demonstrates that the three losses work jointly to provide a better result instead of interfering with each other.

Third, we reach an interesting observation that the relative importance of losses varies much between the structured T2ICompBench prompts and more natural GenAIBench prompts. On T2ICompBench, the specialized binding losses ($\mathcal{L}_{ea}$ for attribute and $\mathcal{L}_{ere}$ for relation) prove most crucial, outperforming the general focus loss $\mathcal{L}_{focus}$ alone, though there is still improvement using $\mathcal{L}_{focus}$ only. However, the behavior shifts notably on GenAIBench, where $\mathcal{L}_{focus}$ provides more substantial improvements than either $\mathcal{L}_{ea}$ or $\mathcal{L}_{ere}$ alone. This finding aligns with the results in Table 1, where Attend-and-Excite (A&E) emerges as the strongest baseline for GenAIBench.

It is important to emphasize that while $\mathcal{L}_{focus}$ drives the most significant gains on GenAIBench, incorporating $\mathcal{L}_{ea}, \mathcal{L}_{ere}$ still yields additional performance improvements. Moreover, on structured benchmarks like T2ICompBench, $\mathcal{L}_{focus}$ alone proves insufficient. These results collectively demonstrate that all three losses play vital though distinct roles in enhancing semantic binding.

Two additional insights emerge from our analysis. First, while $\mathcal{L}_{focus}$ shares some similarity with A&E, its standalone performance surpasses A&E, demonstrating the superiority of our formulation. Secondly, T2ICompBench(Spatial) contains no entity-attribute prompts, making $\mathcal{L}_{focus} + \mathcal{L}_{ere}$ equivalent to the full combination $\mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere}$. In contrast, T2ICompBench(Color) includes some entity-relation-entity prompts, resulting in slight performance differences between $\mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere}$ and $\mathcal{L}_{focus} + \mathcal{L}_{ea}$, a small evidence proving the usefulness of $\mathcal{L}_{ere}$.

6.3 Human Evaluation

To validate that our improvements reflect genuine quality gains rather than metric exploitation, we conduct comprehensive human evaluations across both models. For SD-3, we randomly select 100 output pairs from GenAIBench(Attribute), com-

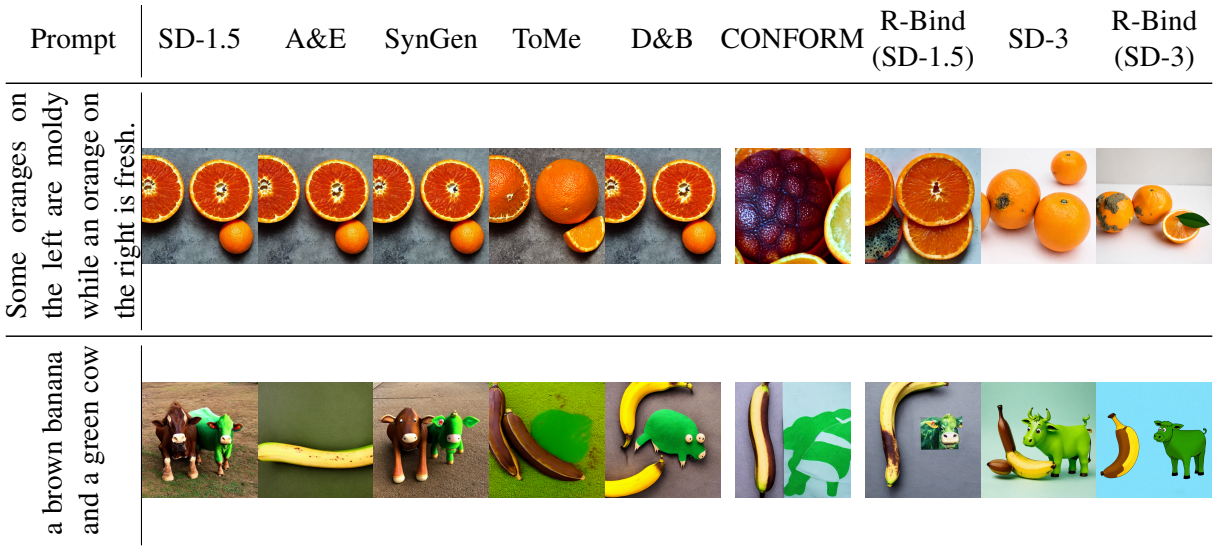

Table 3: Generated results of different methods.

paring base model against SD-3 enhanced with R-Bind. Three independent annotators assessed each pair, selecting the preferred output or marking "draw" for indistinguishable quality through majority voting. We repeat this evaluation on SD-1.5, comparing against two strongest baselines Attend-and-Excite (A&E) and CONFORM. In SD-1.5 “Draw" refers to R-Bind generates one of but not only the preferred results. Details of human evaluation is in Appendix B.

SD-3	R-Bind 0.53	SD-3 0.05	Draw 0.42	- -
SD-1.5	R-Bind 0.61	A&E 0.15	CONFORM 0.08	Draw 0.16

Table 4: Preference rate of generation results on different models and methods.

The human evaluation results in Table 4 demonstrate R-Bind’s consistent superiority. For SD-3, our method produces preferred outputs in over 50% of cases while matching the base model’s quality in 42% of instances ("Draw"). This high Draw rate primarily occurs when SD-3 already generates near-perfect results, leaving minimal room for improvement. Nevertheless, R-Bind still achieves measurable gains in the majority of cases where enhancement is possible.

The SD-1.5 comparisons reveal even more pronounced advantages, with lower Draw rates (indicating more discernible differences) and clear preference for R-Bind over the baselines (A&E and CONFORM). These consistent results across models provide robust evidence that R-Bind’s improvements represent genuine quality enhancements.

6.4 Case Study

Firstly, we present the case after our method is applied in Figure 4 as a comparison with Figure 2.

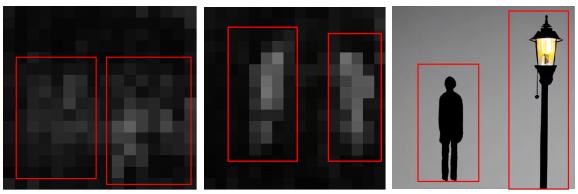

Figure 4: Example of the generation process after R-Bind is applied. The left shows the attention map at the first denoising step, the middle shows the attention map after 10 denoising steps, and the right shows the final generation result.

As can be seen from Figure 4, after 10 denoising steps, the attention map clearly shows two distinct regions attending to the relation "on the left of", each corresponding to one of the entities (man and lamp). This observed behavior matches our intended design, revealing that the method successfully maintains correct attention focus for relation and their associated entities.

We present more cases in Table 3. As can be seen from the cases, the generation results of our method consistently aligns with the text prompt better. For example, in the first line, all methods except ours fail to generate moldy oranges on the left on SD-1.5, and the original SD-3 fails to distinguish moldy oranges on the left and the fresh orange on the right. Our method successfully addresses these problems, showing better performance. More results and analysis can be found in Appendix C.

7 Conclusion

Our work introduces R-Bind, a novel training-free method that improves semantic binding considering entity-relation-entity binding scenarios. By simultaneously optimizing entity-attribute binding and entity-relation-entity binding, our method outperforms all existing baselines on comprehensive benchmarks. R-Bind's effectiveness applies to both UNet-based and DiT-based architectures, demonstrating its practical value for state-of-the-art systems. Rigorous validation through ablation studies, human evaluations, and qualitative analyses further support the effectiveness of R-Bind.

Limitations

Our method is a inference-time optimization method, leading to a higher inference cost compared with base models, yet this is a common problem of all inference-time optimization methods. Also, if the model starts at a really "bad" attention map, our method cannot fix this problem, which is also a common problem of this kind of method.

Ethics Statement

Our method aims at improving alignment between generated image and text prompt, so as long as the text prompt is not harmful, our method will not produce any harmful content. And since we use open-source models and datasets for experiments, the safety of contents in our experiment is generally guaranteed. LLM is used to extract semantics, which is a quite common usage. We conduct human evaluation on the basis of voluntary and each annotator is paid fairly. We also use LLM to assist writing. We use open-source dataset and code following their license and intended uses.

Acknowledgements

This work was supported by Beijing Science and Technology Program (Z231100007423011) and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. 2024. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. *arXiv*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. 2024. Ranni: Taming text-to-image diffusion for accurate instruction following. *Preprint*, arXiv:2311.17002.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, and 1 others. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang, Ming-Ming Cheng, Kai Wang, and Yaxing Wang. 2024a. Token merging for training-free semantic binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37:137646–137672.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. 2024b. Ella: Equip diffusion models with llm for enhanced semantic alignment. *Preprint*, arXiv:2403.05135.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Preprint*, arXiv:2307.06350.

Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. 2024. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *Advances in Neural Information Processing Systems*, 37:76177–76209.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. 2024a. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5290–5301.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22511–22521.

Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. 2024b. Divide bind your attention for improved generative semantic nursing. *Preprint*, arXiv:2307.10864.

Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. 2024. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9005–9014.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.

Gemma Team. 2025. Gemma 3.

Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. 2024. Tokencompose: Text-to-image diffusion with token-level supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8553–8564.

Gaoyang Zhang, Bingtao Fu, Qingnan Fan, Qi Zhang, Runxing Liu, Hong Gu, Huaqi Zhang, and Xinguo Liu. 2024. Compass: Enhancing spatial understanding in text-to-image diffusion models. *Preprint*, arXiv:2412.13195.

A Details of Our Method

A.1 Multiple Tokens

As is mentioned, given an entity-attribute pair (e, a), the entity may contain multiple tokens $e = (e_{t1}, ..., e_{tk})$, and the attribute may also contain multiple tokens $a = (a_{t1}, ..., a_{tl})$. The formulation mentioned in Section 4 is a conceptual simplified representation, and we would like to present details on how to handle these as follows:

Firstly, these multiple entity tokens $e = (e_{t1}, ..., e_{tk})$ jointly represent a certain entity, so we average the attention map of all these tokens, in this case we have $\tilde{A}[e] = \frac{1}{k}\sum_{i=1}^{k} A[e_{ti}]$.

Secondly, these attribute tokens may present different attributes. For example, consider a prompt "a brown fat cat", the attribute tokens are (brown, fat). Therefore, we would like to optimize the worst semantic binding of all attributes. Formally, we have:

$$D(\tilde{A}[e], A[a]) = \max_i D(\tilde{A}[e], A[a_{ti}]) \tag{9}$$

However, when calculating $K((e_i, a_i), (e_j, a_j))$, separating all these tokens can be rather complicated. Therefore, when calculating K, we also average the attention map of all attribute tokens, which is $\tilde{A}[a] = \frac{1}{l}\sum_{i=1}^{l} A[a_{ti}]$.

For clearer notation, we use:

$$\tilde{D}_t(t_1, t_2) = D(\tilde{A}[t_1], \tilde{A}[t_2]) \tag{10}$$

Thus we have:

$$\begin{aligned} K((e_i, a_i), (e_j, a_j)) = \tilde{D}_t(e_i, e_j) + \\ \tilde{D}_t(e_i, a_j) + \tilde{D}_t(e_j, a_i) + \tilde{D}_t(a_i, a_j) \end{aligned} \tag{11}$$

The calculation of K follows the same requirement as mentioned in Section 4.

The final $\mathcal{L}_{ea}$ considering multiple tokens is represented as:

$$\begin{aligned} \mathcal{L}_{ea} = \sum_{(e_i, a_i) \in S_{ea}} [D(\tilde{A}[e], A[a]) - \\ \frac{1}{|Z|} \sum_{(e_j, a_j) \in S_{ea}} K((e_i, a_i), (e_j, a_j))] \end{aligned} \tag{12}$$

Similarly, given an entity-relation-entity triplet (e^1, r, e^2), the entity may contain multiple tokens,

which we deal with as before. The relation may also contain multiple tokens $(r_{t1}, ..., r_{tu})$. Similarly, we would like to optimize the worst semantic binding, which is:

$$\max_i(D(\tilde{A}[e^1], A[r_{ti}]) + D(A[r_{ti}], \tilde{A}[e^2])) \quad (13)$$

Denote $\tilde{A}[r] = \frac{1}{u}\sum_{i=1}^{u} A[r_{ti}]$, we have:

$$K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2)) = \tilde{D}_t(e_i^1, r_j) + \tilde{D}_t(e_i^2, r_j) + \tilde{D}_t(e_j^1, r_i) + \tilde{D}_t(e_j^2, r_i) + \tilde{D}_t(e_i^1, e_j^1) + \tilde{D}_t(e_i^2, e_j^2) \quad (14)$$

This calculation of K also follows the same requirement as mentioned in Section 4.

So the final $\mathcal{L}_{ere}$ considering multiple tokens is:

$$\mathcal{L}_{ere} = \sum_{(e_i^1, r_i, e_i^2)} \max_i(D(\tilde{A}[e^1], A[r_{ti}]) + D(A[r_{ti}], \tilde{A}[e^2])) - \min(\tilde{D}_t(e_i^1, e_i^2), \frac{1}{|Z|}\sum_{(e_j^1, r_j, e_j^2)} K((e_i^1, r_i, e_i^2), (e_j^1, r_j, e_j^2))) \quad (15)$$

A.2 Algorithm

To provide a more comprehensive understanding of our algorithm, we present a pseudo code in Algorithm 1:

Algorithm 1 Denoising with R-Bind

Input total denoising steps T, noise prior x_T, denoising network ϵ_θ, text prompt p, noise scheduler F, guidance scale $\tilde{w}$, text encoder E, R-Bind step threshold T_0, optimization steps T_1, optimization step size α.
Get text embedding $c = E(p)$
Extract semantics S_e, S_{ea}, S_{ere} from p.
for $t = T, ..., 1$ **do**
  if $t \leq T_0$ **then**
    for $s = 1, ..., T_1$ **do**
      Run forward $\epsilon_\theta(x_t, t, c)$ to achieve attention map A
      Calculate $\mathcal{L}_{focus}, \mathcal{L}_{ea}, \mathcal{L}_{ere}$ using A, S_e, S_{ea}, S_{ere}
      $\mathcal{L} = \mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere}$
      Update $x_t \leftarrow x_t - \alpha\frac{\partial\mathcal{L}}{\partial x_t}$
    end for
  end if
  Predict $\epsilon_\theta(x_t, t, c), \epsilon_\theta(x_t, t, \phi)$
  Classifier-Free Guidance: $z_t \leftarrow \epsilon_\theta(x_t, t, \phi) + \tilde{w}(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \phi))$
  Denoising Step: $x_{t-1} \leftarrow F(x_t, z_t, t)$
end for
Output denoising result x_0

B Details of Experiment Setup

B.1 Models, Benchmarks and Hyper-Parameters

We use SD-3 and SD-1.5 using their default hyper-parameters. The checkpoints and hyper-parameter used are as follows:

Model Name	Checkpoint	T	w
SD-3	SD-3-Medium [1]	28	7.0
SD-1.5	SD-1.5 [2]	50	7.5

Table 5: Details of our inference hyper-parameter.

For our method, we perform optimization in the first 50% steps in the inference following previous practice (Rassin et al., 2023), corresponding to $T_0 = \frac{T}{2}$. We perform optimization twice per denoising step, corresponding to $T_1 = 2$. The optimization step size is set to $\alpha = 6$ for SD-1.5 and $\alpha = 8$ for SD-3 since SD-3 is a larger model. Also, we observe that the optimization process of SD-3 is sometimes not quite stable, so we apply an additional adaptive operation on SD-3, which is $\alpha_t = e^{\mathcal{L}}\alpha$ at each optimization step t. Note that this operation is not necessary and our method generally works fine without it. This also suggest a future direction for dynamic tuning of the optimization step size in test-time optimization methods, which we leave for future study. We present an experiment on the influence of step size in Appendix C.3.

We also list benchmark statistics in Table 6:

T2ICompBench (Color)	T2ICompBench (Spatial)	GenAIBench (Attribute)	GenAIBench (Spatial)
300	300	1000 [3]	831

Table 6: Benchmark statistics.

[1] https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers

[2] https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5

[3] We randomly sampled 1000 prompts belonging to this category.

B.2 Details of Semantic Extraction

We utilize Gemma-3-27B (Team, 2025) to performan semantic extraction since it is a powerful LLM. Note that this semantic extraction is a **text-only** task.

The Semantic Extraction process involves two stages: parsing the input prompt to identify and categorize semantic information (entities, entity-attribute, and entity-relation-entity), and token mapping of these elements according to the diffusion model's text tokenizer to produce the token sequences required for attention map manipulation. This dual-stage approach ensures that our binding losses operate on precisely the same textual representations used by the diffusion model's cross-attention mechanisms during image generation.

In our work, both steps are conducted by the LLM. For the first step, the prompt used is as follows:

Correctness Verification To validate the reliability of our semantic extraction pipeline, we performed manual verification on 100 randomly sampled prompts from GenAIBench(Spatial), finding 93% exact match accuracy between the LLM's extraction results and ground truth annotations. This high accuracy confirms the LLM's effectiveness for semantic extraction in our context. Also, for the structured prompts in T2ICompBench(Color) and T2ICompBench(Spatial), the structure of the prompts guarantees perfect (100%) extraction accuracy.

Discussion of LLM Usage While many prior works rely on custom rule-based parsers for semantic extraction, such approaches face significant challenges in handling the full complexity of real-world prompts, particularly when dealing with intricate relations and tokenization (e.g., splitting one word into multiple tokens). To ensure robust generalization, we instead employ an LLM (Gemma-3-27B) as our semantic extractor. Crucially, our experiments confirm that the performance gains stem from our novel binding framework rather than the use of LLM. In fact, CONFORM (Meral et al., 2024) and D&B (Li et al., 2024b) do not open-source the parser their used, so we also equip these methods with the same LLM (Gemma-3-27B) when testing them. However they still underperform compared with our method R-Bind, as can be seen from Table 1. Also, on T2ICompBench, where LLM and parser extractions yield identical results, our method maintains clear superiority. These results demonstrate that our semantic binding methods, not the use of LLM, drive the observed improvements. We would also like to note that, our selected LLM is not the only possible way of performing semantic extraction. Other strategies, such as smaller LLMs, better-designed rule-based parsers, or even human annotation (which is quite common for practical usage, as many works require additional human input instead of providing a parser).

B.3 Details of Human Evaluation

We ask three human annotators to rank the images based on the alignment between image and text prompt only. The inter-annotator agreement is 0.93. All annotators are college students and are capable and responsible of conducting this task. A simplified evaluation criteria is shown as follows:

Human Evaluation Criteria

Please select the image that aligns with the text best from the given images. You can select more than one image if you believe the consistency between your selected images and the text is comparable. The consistency between image and text indicates whether the image faithfully describes the contents mentioned in the text.

Our annotation protocol applies majority voting to achieve the final result. The images selected by most annotators are viewed as the winner. If R-Bind and another baseline are selected the same times, we label this a "Draw". There are no cases where both baselines are selected the same times.

C More Results and Analysis

C.1 Discussion on Efficiency

Inference-time optimization bear a natural worry of efficiency. We admit that our method does make inference slower, yet we argue that this efficiency decrease is acceptable and not significantly beyond other inference-time optimization methods. We present the efficiency comparison in Table 7.

As can be seen from Table 7, our method, though a lot slower than base model, bear similar inference time with most other baseline methods, indicating that our method does not bear severe efficiency problem compared with other baseline methods.

Extraction Prompt

[System Prompt]: You are a helpful assistant good at extracting information from complex text. You will be given a text and your task is to extract three types of information from the given text. The three types of information are:
Entity Information, which is the entities mentioned in the text.
Entity-Attribute information, which is a tuple containing an entity and the attribute describing it.
Entity-Relation-Entity information, which is a tuple containing two entities and the relation between them.
Please extract the three types of information from the given text. You should output them as: Entity Information: [Entity information], Entity-Attribute Information: [entity-attribute information], Entity-Relation-Entity Information: [entity-relation-entity information]. If there is no such information belonging to such category, you should output [None].
If there are pronouns in the text, you should correctly replace them with the corresponding entities in the extracted information.
Entities with no attribute should not appear in Entity-Attribute information, same for Entity-Relation-Entity information.
Do not mix entity-attribute information and entity-relation-entity information. If the attribute of an entity is a verb, please check whether it is entity-relation-entity information.
Do not miss any entity-attribute information and entity-relation-entity information. You should output all reasonable extracted information.
[In-Context Examples]
[User Prompt]: The provided text prompt is {text}.
[Model Output]:

Token Matching Prompt

[System Prompt]: You are a helpful assistant good at matching token id with extracted information from a complex text.
You will be given the text and extracted information from the text and a corresponding token list. Your task is to replace the information in the extracted information with correct token id using the token list.
There are three types of information: Entity Information, which describes the entities mentioned in the text.
Entity-Attribute information, which is a tuple containing an entity and the attribute describing it.
Entity-Relation-Entity information, which is a tuple containing two entities and the relation between them.
You should output them as: Entity Information: [Entity information], Entity-Attribute Information: [entity-attribute information], Entity-Relation-Entity Information: [entity-relation-entity information]. If there is no such information belonging to such category, you should output [None].
You should use token ids to represent, entity, attribute and relation as your final output. Each entity, attribute, relation can be represented using one or multiple token ids.
The Entity-Attribute information should be represented as: (token ids of entity, token ids of attribute). The Entity-Relation-Entity information should be represented as: (token id of entity 1, token id of relation, token id of entity 2). The Entity Information should be represented as: (token ids of entites). [None] should not appear in any certain information tuple.
[In-Context Examples]
[User Prompt]: The provided text prompt is {text}. The extracted information is {information}. The token list is {tokens}.
[Model Output]:

Method Name	Seconds Per Image
Base Model	2.12
A&E	8.31
SynGen	5.79
ToMe	6.78
D&B	13.75
CONFORM	9.71
R-Bind(Ours)	11.69

Table 7: Average seconds required for each pipeline. Semantic extraction is not counted as part of the time since there can be multiple possible ways of performing semantic extraction, such as using parsers, LLMs of different sizes or even human annotations.

For better mitigating efficiency problem, we conduct another experiment using $T_0 = \frac{T}{4}$ on SD-1.5. The results are shown in Table 8.

Method Name	T2ICompBench(Spatial)	SPI
SynGen	10.9	5.79
R-Bind ($T_0 = \frac{T}{2}$)	15.6	11.69
R-Bind ($T_0 = \frac{T}{4}$)	14.0	7.78

Table 8: Results under different optimization steps T_0.

As can be seen from the results, setting $T_0 = \frac{T}{4}$ further improves efficiency with acceptable performance drop, still outperforming baseline method SynGen, indicating a better potential for efficiency and performance tradeoff.

C.2 Discussion on Loss Weighing

Our proposed final loss function is shown as:

$$\mathcal{L} = \mathcal{L}_{focus} + \mathcal{L}_{ea} + \mathcal{L}_{ere} \tag{16}$$

which provides an equal weighing of the three components. However, to delve deeper, the loss can be written as:

$$\mathcal{L} = \alpha\mathcal{L}_{focus} + \beta\mathcal{L}_{ea} + \gamma\mathcal{L}_{ere} \tag{17}$$

We conduct another experiment on SD-1.5 to explore whether different loss weighing have a great influence on our method. The result is shown in Table 9.

As the results demonstrate, selecting $\alpha = \beta = \gamma = 1$(equal weighing) already yields highly competitive results. The marginal performance changes observed with different weightings confirm that our method is robust and not overly sensitive to the loss weight configuration. This validates our choice of using equal weights as a practical and effective default setting.

α	β	γ	Color	Spatial
1	1	1	**68.8**	15.6
1.5	1	1	67.8	13.0
1	1.5	1	67.3	15.6
1	1	1.5	**68.8**	**15.9**

Table 9: Results of different loss weighing. “Color" and “Spatial" refer to corresponding splits of T2ICompBench.

C.3 Discussion on Step Size

The optimization step size α in Equation 1 is an interesting hyper-parameter. We analyze its influence on SD-1.5. The results are shown in Table.

α	Color	Spatial
4	65.4	13.7
6	68.8	15.6
8	68.6	15.8

Table 10: Results of different optimization step sizes. “Color" and “Spatial" refer to corresponding splits of T2ICompBench.

As the results show, modifying α does not have a significant impact on our method’s performance, as long as α is within a reasonable range (e.g., from 6 to 8). However, α should not be too small. Also, even at a suboptimal small value, our method still provides competitive results when compared to other baselines in Table 1, further highlighting its effectiveness and robustness.

C.4 Failure Case Analysis

No method is perfect and it is natural for any method to fail on some cases. Here we would like to analyze why our method fails on a certain case. The failure case is shown in Figure 5.

We attribute the failure of this case to the bad initial attention map. As can be seen from the attention map of “girl" at $t = T$, which is the first denoising step, the attention map is rather scattered and has no focus on the entity “girl" itself. After our method is applied 20 steps, the attention map is still scattered, though slightly better than the original. As a result, the model actually has no

Table 11: Generation results on different models and methods.

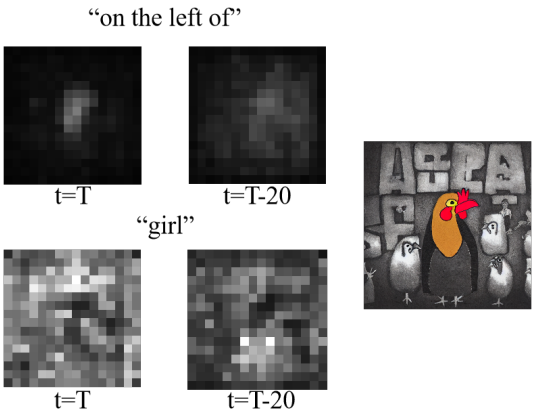

Figure 5: A failure case using R-Bind. The left images are attention maps of corresponding tokens at certain steps. The right is the generation result. The prompt is “a chicken of on the left of a girl".

idea how to generate the entity “girl", let alone the relation “on the left of". This example shows that if the original attention map is much flawed, our method, though still able to improve the attention map, fails to completely address the problem since it is just an inference-time optimization method.

C.5 More Case Study

We present more generated examples in Table 11.